Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning

Volume 42:

https://sites.google.com/site/hepml14

Higgs challenge
When HEP meets ML
the HiggsML challenge
www.kaggle.com/competitions
ATLAS EXPERIMENT
LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE
Inria INVENTEURS DU MONDE NUMÉRIQUE
CERN
Google
The challenge is organised by the ATLAS collaboration, the LAL (AppStat and Atlas groups) and INRIA-Saclay, in partnership with CERN and Google

Edited by: Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl and David Rousseau

Preface

This volume gathers papers contributed to the NIPS 2014 workshop in High Energy Physics and Machine Learning (HEPML). It includes papers from a unique experiment in high energy physics: using the power of the "crowd" to help solving difficult physics problems. This experiment took the form of a data science challenge organized in 2014 [1] and whose results were discussed at the workshop

To discover and study new particles, characterize the properties of matter and test new theories of physics, experiments are conducted on the Large Hadron Collider (LHC) at CERN where proton bunches are accelerated on a circular trajectory in both directions. When these bunches cross in the a detector, some of the protons collide, producing hundreds of millions of proton-proton collisions per second. Up to hundreds of particles resulting from each bunch crossing (called an *event*) are detected by sensors, producing a sparse vector of about a hundred thousand dimensions (roughly corresponding to an image or speech signal in classical machine learning applications). This is where machine learning can enter into play: recognize the particle signatures left in the detectors.

The volume of data produced in experiments conducted at CERN is mind boggling, approaching petabytes per seconds. Not all of the data can be saved. Real-time multi-stage cascade classifiers (called the *trigger*) discard most of the uninteresting events. Vladimir V. Gligorov gives an introduction in this volume to methods used at CERN to perform real-time data analysis at the LHC. The selected events (roughly four hundred per second) are then written on disks by a large CPU farm, producing petabytes of data per year. The saved events still, in large majority, represent known processes: they are mostly produced by known processes having been discovered in previous generations of experiments.

The goal of the challenge was to improve the offline analysis (the next stage) to find regions in feature space in which there is a significant excess of *signal* events compared to what known background processes can explain. Once the region has been fixed, a statistical (counting) test is applied to determine the significance of the excess. If the probability that the excess has been produced by only background processes falls below a limit (usually corresponding to a 5σ confidence or a pvalue of the order of 10^{-7}) a new hypothesized particle is deemed to be discovered. The problem of the Challenge is therefore a new problem in machine learning that we call "learning to discover". This superficially resembles a classification problem (separate signals from backgrounds) but it has very new distinctive difficulties tied to the very particular objective to be optimized: the power of a statistical test. This was exemplified in the Challenge on a specific question: improving the characterization of the Higgs Boson, by studying a particular decay channel called Higgs to Tau Tau, whereby the famous Higgs particle, whose discovery was recently claimed jointly by the ATLAS and CMS experiments at CERN, decays into a fermion pair, namely tau leptons. This complements the original discovery of the Higgs boson, which was first seen in decay channels involving

[1] https://www.kaggle.com/c/higgs-boson

boson pairs. Due in part to the publicity around the discovery of the Higgs boson in popular science and the recent Nobel prize, the participation was overwhelming with more than 1700 teams. Machine learning specialists, physicists and students, submitted more than 30.000 solutions, making it one of the most popular data science challenges to date.

Data science challenges have proved to be efficient and cost-effective ways to quickly bring to new application domains effective solutions engineered by the best data scientists. In many domains, recurrent events are organized yearly to stimulate the scientific community. Examples include CASP, CAMDA, and DREAM in genomics and proteomics, TREK in text processing, TREK VID, IMAGENET, BRATS and CHALEARN LAP in image and video processing, ROBOCUP in robotics, CHALEARN causal discovery challenges [2], etc. Challenges are posted on-line on platforms such as Kaggle and TopCoders. New ground-breaking algorithms have been developed through that effort, with applications in many domains including genomics, drug discovery, epidemiology, and neuroscience. While in any particular domain, experts may take years to develop good solutions to difficult problems, it is not uncommon that submitting the problem to the public in the form of a challenge significantly improves performance.

The Challenge was an unambiguous demonstration of the the potential of the machine learning crowd. The objective function, representing the significance of discovery of a new particle, of the top ten participants is in the 3.76-3.80σ range, while the untuned baseline benchmark (based on the TMVA software[3] widely used in High Energy physics) was $\sim 3.20\sigma$, the MultiBoost benchmark[4] $\sim 3.40\sigma$ and the tuned version of TMVA ranked only 782 at $\sim 3.50\sigma$. More details on the challenge are given in the paper summarizing the results.

The data were released on the CERN Open Data Portal[5] after the end of the challenge. This unprecedented disclosure of precious data belonging to the ATLAS collaboration, highlights the importance of the learning to discover task. The dataset, and the subject of the challenge correspond to particular physical process, the $H \rightarrow \tau^+\tau^-$channel. However, the methodology is fully generic to the discovery of a new particle, and could generalize to other discovery settings.

The outcome of the challenge was interesting to high energy physicists in several respects, illustrated by the papers of this volume contributed by some of the top ranking participants. The winning method of Gábor Melis, which used a Deep Learning method, was computationally expensive but beats significantly the runner up. The success of the winner is attributable in part to the very careful way in which he conducted cross-validation to avoid overfitting. A special "High Energy Physics meets Machine Learning" award was given to team Crowwork (Tianqi Chen and Tong He) who had a slightly lower score but provided a method called XGboost that is a good compromise between performance and simplicity, which could improve tools currently used in high-energy physics.

The paper of Gábor Melis is also relevant to the promise that deep learning methods can save on human effort by learning internal representation in place of feature engineering, that was not held in the challenge in two respects. Firstly, the optimization of the architecture and hyper-parameters of the winning entry required skilled human expertise

[2] http://www.causality.inf.ethz.ch/challenges.php

[3] http://tmva.sourceforge.net

[4] http://higgsml.lal.in2p3.fr/software/multiboost/

[5] http://opendata.cern.ch/

(and month of GPU computing). Secondly, the performance dropped by 13% when only the most primitive features were used, compared to using derived features incorporating knowledge of the underlying physics. This volume also includes another paper by Sadowski et al. providing a rather extensive benchmark of deep and shallow networks on various high energy physics tasks. They also show a disappointing behavior concerning learning internal representation for the Higgs in tau tau task, with a 6% performance drop , although the results are more encouraging for other high energy physics cases. The setting of their paper is not directly comparable to that of the challenge because Sadowski et al. used a larger and simpler dataset of millions of events. However, the ratio of the gain to sample size indicates that learning representations is extremely data demanding for this task.

From the machine learning research point of view, one of the key difficulties was to optimize the very peculiar objective function called Approximate Median Significance (AMS) devised by the physicists to identify and test the significance of a region in feature space where the number of signals exceed the number that the distribution of background particles alone could explain. The AMS is particularly nasty since it is discontinuous, non differentiable, non additive (the overall AMS is not the sum of individual contributions of the samples), and it uses sample weights available only for training. This added with the fact that the number of signals is orders of magnitude smaller that the number of backgrounds made the problem particularly difficult. Most participants proceeded in two stages: first train a regular learning machine with a surrogate objective function better behaved computationally than the AMS, such as the cross-entropy, then modify the bias value of the classifier to optimize the AMS using cross-validation. The paper of Wojciech Kotlowski puts a theoretical framework around this approach to prove its consistency. The paper of Roberto Díaz-Morales and Ángel Navia-Vázquez introduces a surrogate objective function incorporating the weight of the training sample: the weighted AUC. This volume includes a contribution by Lester Mackey, Jordan Bryan, and Man Yue Mo who use a variational method to iteratively optimize an approximation of the AMS having desirable computational properties.

The workshop generated an intense discussion. Among the topics addressed, **increasing the impact of the contribution of machine learning** was of obvious interest to the audience, who tried to understand where the bottleneck of the entire analysis pipeline presently lies. It appeared by consulting with the physicists that, with the improvement of detectors, the amount of data being recorded is ever increasing, thus requiring to make quicker and quicker decisions about which event of potential interest to record. Presently, hardware triggers designed with expert knowledge take care of this, but one can envision that machine learning algorithms, either implemented in hardware or in software, could make it to the front end of the pipeline.

Another topic concerned the **validity of the machine learning solutions trained on simulated data**. One key issue the audience tried to grasp was whether it would be possible to train machine learning models with real data to limit the *systematic* error. The problem with real data however is that, prior to a discovery, only background events can reliably be identified and labeled in real data. The identification of the signature of signal events in real data is precisely the problem that needs to be solved, making the problem of learning to discover close to a novelty detection problem. Using simulated data to train is logical because novel physics theory produces data generating models, which can provide simulated signatures of signals and backgrounds. Hence the role of machine learning is to create pattern recognition models, which invert the data generating models. But of course, the simulators have many tunable parameters and the learning machines may faithfully learn intricate details of the simulated distribution, which are not relevant to the real data. The issue of mixing real and artificial data either for training or testing was raised. This touches upon of hot topics in machine learning: semi-supervised learning and transfer learning.

The participants were also concerned that **the dataset from the official ATLAS simulator provided for the challenge was too small** to effectively train learning machines. Indeed, a lot of challenge participants reported the extreme difficulty not to overfit the AMS objective function. The winners succeeded in avoiding overfitting by making an effective use of regularization and cross-validation. However, more powerful learning machines could be trained with more data, such as larger and deeper neural networks. Comparisons were made with the paper of Sadowski et al., which used an order of magnitude more data. The dataset however was generated with a simpler simulator, not accounting for many imperfections of the detectors, so the learning machines are not directly usable. The question was raised whether it would be possible to somehow combine datasets. This may allow deep learning methods and other methods to learn data representations from primitive features, eliminating the need to incorporate human knowledge in the form of derived features.

The validity of potential discoveries made with the test obtained from machine learning was also challenged because of the problem of multiple testing (if several tests are tried by multiple researchers and only the results of one of them is reported, without taking into account the number of attempts made). This problem has been scrutinized by HEP for years under the name of the *look elsewhere effect*. Whether Machine Learning could contribute, *e.g.* in the form of performance bounds taking into account the number of trials, is an open question.

Another concern was that, in an attempt to simplify the problem, the objective function used in the challenge (the AMS) is not exactly what physicists use. The state-of-the-art methodology involves testing differences in the whole distribution of signals and backgrounds using binning. This opens up a very interesting direction of research on how to build effective surrogate objective function, which can be optimized by learning and addresses well the true objective.

The conclusion of the event was: do it again! Several new challenges are under study, including a challenge to improve tracking *i.e.* tracing the trajectory of the particles faster and more reliably.

June 2015

The Editorial Team:

Glen Cowan
Physics Department, Royal Holloway, University of London, UK G.Cowan@rhul.ac.uk

Cécile Germain
LRI, University Paris-Sud & CNRS & INRIA, France Cecile.Germain@lri.fr

Isabelle Guyon
ChaLearn and ClopiNet, Berkeley, California, USA
guyon@chalearn.org

Balázs Kégl
LAL, IN2P3/CNRS & TA0, INRIA & LRI, University Paris-Sud & CNRS, France
kegl@lal.in2p3.fr

David Rousseau
LAL, IN2P3/CNRS & University Paris-Sud, France rousseau@lal.in2p3.fr